A Hybrid Approach to Semi-Supervised Named Entity Recognition in Health, Safety and Environment Reports

Yunita Sari[1], M. Fadzil Hassan[2], Norshuhani Zamin[3]
Department of Information and Computer Sciences
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia
sari.yunitata@yahoo.com[1], mfadzil_hassan@petronas.com.my[2], norshuhani@petronas.com.my[3]

***Abstract*—In the last few years, text mining have become the area of interests in Natural Language Processing (NLP). They share a similar idea i.e. to extract important facts from unstructured text which later help to populate database entries. Name Entity Recognition (NER) is one of the main task needed to develop text mining systems in which it is used to identify and classify entities in the text into predefined categories such as the names of persons, organizations, locations, dates, times, quantities, monetary values, percentages, etc. This paper focuses on studying the optimum solution to perform NER. To achieve our target, Health Safety and Environment (HSE) reports available from the Universiti Teknologi PETRONAS (UTP) are chosen as the case study. The UTP's HSE reports are the investigation reports which contain the information on incidents and accidents occurred during the daily operations. Many algorithms have been reported for NER ranging from simple statistical methods to advanced Natural language Processing (NLP) methods. This paper describes the possibility to apply Link Grammar (LG) and Basilisk Algorithm in NER.**

Keywords: Name Entity Recognition, Health and Safety Environment, Link Grammar, Basilisk Algorithm.

I. INTRODUCTION

Named Entity Recognition (NER) was one of the main topics of interests during the sixth Message Understanding Conference (MUC-6) [1]. In the NER, systems are required to recognize only the named entities occurring in a text. In the UTP's HSE reports domain, the likely entities are date of accident, location of accident, time of accident, accident type, instrument involved and effect of the accident. The goal of NER is to semantically recognize and identify the occurrences of some predefined phrase types in an annotated text [2].

As NER is a subtask of text mining, limited number of research was found on NER. Although more efforts have been made to focus on developing text mining systems such as Information Retrieval, Information Extraction and Question Answering but we believe that a NER plays an important role to increase an accuracy of such systems. Most of the existing NER algorithms are domain dependent. Generally, different domain will have different patterns in the structure of the text. In this case, a set of specific pattern rules are needed based from the nature of UTP's HSE reports. For instance, date identification used in accident report and date used in e-mail text. Though both of them have the same goal, but the syntactic analysis to create an extraction pattern is different.

In this paper, we are proposing an NER algorithm using Link Grammar (LG) [6] and Basilisk bootstrapping Algorithm [7] for extracting entities and building semantic lexicon based on UTP's HSE report. To the best of our knowledge, within the limited literatures available on NER, LG and Basilisk have never been attempted to recognize and classify entities in any HSE reports. Hence, we can foresee the challenges where the availability of a complete dictionary or semantic lexicon for HSE domain will be almost impossible.

II. OBJECTIVE

The objective of this paper is to introduce an NER system architecture that uses LG parser and Basilisk bootstrapping Algorithm for recognizing some entity names in the UTP's HSE reports and building semantic lexicon or dictionary of HSE terms.

III. RELATED WORKS

A. General NER

NER is an ongoing research. A thesis in [2] briefly discussed NER efforts made between 1991 and 2006. The direct goal of some task related with NER isn't always to recognize the named things from document. An NER research known as Person Name Entity Recognition for Arabic (PERA) to recognize and extract person name for the Arabic language using a rule based approach was described in [3]. The system has two major components, the gazetteers and the grammar. Techniques used for collecting total of 472617 entries to build dictionaries. The gazetteer plays as fixed static dictionary that will identify person name in input text. Combination of regular extraction patterns were used to create grammar that performs recognition and extraction. And the last, filtration mechanism is used to separate invalid person names. An integrated software application FAST ESP was used to implement PERA system. The system achieved 85.5% precision and 89% recall for the total 46 data sets.

HaSIE [4] is a complete Information Extraction system developed using the General Architecture for Text

978-0-7695-3591-3/09 $25.00 © 2009 IEEE
DOI 10.1109/ICFCC.2009.52

Engineering (GATE) [5] architecture. HaSIE aims at producing a health and safety summary from annual company reports. A set of NER module of information extraction component included in GATE called A Nearly-New IE system (ANNIE) was used by HaSIE. Tokeniser, sentence splitter, part of speech tagger are directly adopted from ANNIE. Two another components, gazetteer and semantic tagger were modified to adjust the needs of the system. ANNIE mainly used finite state algorithm to perform all process. HaSIE achieved 80% precision and 83% recall based on ten documents data sets.

A personal name disambiguation system is built in [8]. The system identifies the correct reference of a given designator. Another research called Acronym Identification system was attempted in [9]. It identifies an acronym in a given document. For instance the system will identify IBM as the acronyms of "International Business Machine".

A. Mansouri et al. [12] classify these researches into three classes, Hand-made Ruled-based NER, Machine Learning-based NER and Hybrid NER. Hand- made Ruled-based using handcraft rules set to extract entities. This technique shows good result for specific domain but need a high cost to maintain the rules.

D. Nadeau [2] grouped Machine Learning-based NER technique into three categories, Supervised Learning (SL), Semi-Supervised Learning (SSL) and Unsupervised Learning (UL). Hidden Markov Model, Decision Trees, Maximum Entropy Models, Support Vector Machine and Conditional Random Field are variants of SL. SL approach usually achieve good result if amount of the training data is large. Decision Trees for name entity recognition and classification used in [10]. The system uses the C4.5 algorithm. An overwhelming performance results are discovered against the manual system where 89.6% recall and 86.6% precision on the performance of the representative classifier, while a manually constructed system on the same data achieved 69.25% recall and 83.42% precision.

B. Semi-Supervised Learning

As mentioned before, SSL is Machine Learning-based NER technique. Different with SL technique, SSL doesn't need a large amount of training data. Yangarbear et al., in their research [13] explains some reasons why learning method is chosen instead of relying on the complete gazetteer. A fact that a complete dictionary is difficult to obtain and new names periodically enter into the literature, become a main reason. SSL just needs a set of seeds to start the learning process. With this small data, system will identify some contextual clues and reapplied this process into another example data found. As the result, repeating process will create a large number of data and increase the size of the semantic lexicon.

The advantage of this technique is it can be easily port into different domain. Bootstrapping algorithm called Basilisk [7] is one of research focused on SSL technique which was demonstrated on the terrorism attack domain. This algorithm needs unannotated text and small set of data as the input. Basilisk Algorithm used extraction pattern context to identify new word that match with specific category and gather this new words to build semantic lexicon. As the pattern extractor, AutoSlog system [11] is used. Every noun phrase will be extracted into three syntactical roles: subject, direct object and prepositional object. In their experiment, they used Basilisk Algorithm to identify six semantic categories: building, event, human, location, time and weapon. Over hundreds of words, Basilisk achieved 79.5% accuracy for humans and 53.2% accuracy for locations.

C. Link Grammar

LG is proposed as the pattern extractor. LG is one of the syntactic parser for English text. It provides a set of labeled links connecting pairs of words and also produces a "constituent" representation of a sentence (showing noun phrase, verb phrase, etc). LG is often used in information extraction works.

A research by [14] tried to identify events in input text and extract information about them. First, they identified all sentences in input text where the verb represents an action. Then using structure given by LG, a set of rule is made to identify the subject and object. Surprisingly, the result of this works was tested on the Reuters corpus achieved close to 100% precision and recall.

IV. METHODOLOGY AND APPROACH

In this research, we propose an NER system architecture that will recognize entities in UTP's HSE reports. Entities that will be recognized are crucial attributes of an accident like date of accident, location of accident, time of accident, accident type, instrument involved and effect of the accident. Fig. 1 illustrates the system architecture.

Generally, the system architecture can be divided into three processes as highlighted and labeled in the Fig. 1 as three different dashed boxes.

Process 1: Preprocessing Algorithm

Preprocessing Algorithm is the step for preparing the input, in this case the UTP's HSE reports. As indicated in Fig. 1, there are two sub processes involve. The first sub process is the application of the Text Cleaning Algorithm where all images and blanks spaces in unstructured texts will be removed as they are considered noisy elements. Clean text will be produced from this sub process. The second sub process is the application of Parsing algorithm. Here is where the LG parser is applied to parse and annotate the text with lexical category and linkages. Fig. 2 shows an example of LG parser.

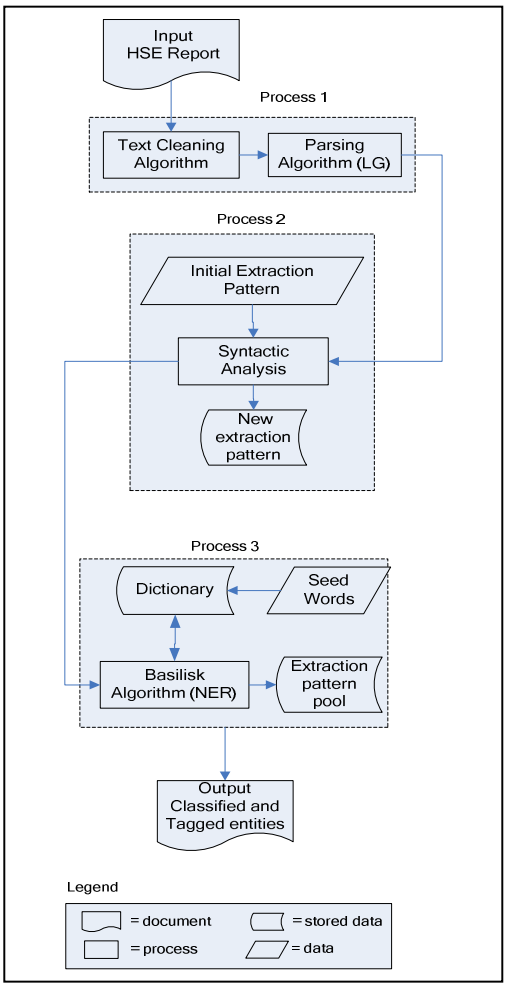


Figure 1. System Architecture

In the sample output of an LG parser in the Fig. 2, the sentence will be parsed into words and each pair of words that relate to each other are connected by specific connector. In the example, word "the" and "accident" are connected by `Ds` connector. This connector connects determiner to single noun. The advantage of LG parser is it extremely reduces the effort to annotate all examples. The linkages give the information how each words are semantically and syntactically associated to each other.

Process 2: Syntactic Analysis

This process is motivated by a work in [7]. Here, the extraction patterns will be created based from a specific pre-determined pattern's frequency. Process of creating extraction pattern begins with adding some initial extraction patterns into the system. From manual analysis over the accident report, extraction patterns will be created. For instance, when we want to extract an accident location and the phrase "the accident happened in" is frequently appear in most of the reports, we can consider "*the accident happened in <location>*" as a possible extraction pattern.

With the aid of LG, better extraction pattern can be created by considering the linkages between words. For example in Fig. 2, there is `Js` connector between preposition "in" and noun phrase "The Chancellor Hall". `Js` connect preposition with their object. The extraction pattern result will be more effective because it also consider syntactical structure between words. The new extraction pattern will be stored in the new extraction pattern.

```
                                +--------Js--------+
                                |   +-------DG------+
     +---Ds--+----Ss-----+--MVp-+   |       +---G---+
     |       |          |      |   |       |       |
   the accident.n happened.v in the Chancellor Hall

Constituent tree:
(S (NP The accident)
   (VP happened
       (PP in
           (NP the Chancellor Hall))))
```

Figure 2. Link Grammar Parser

Process 3: Named Entity Recognition and Bootstrapping

Named Entity Recognition and Bootstrapping are the processes to identify and add new entities found into the dictionary list. Basilisk Algorithm that was implemented by [7] will be applied. In their research, they used this algorithm to learn new word or entity that has similar extraction pattern context with extraction pattern generated by the pattern extractor.

With semi-supervised learning, the proposed hybrid architecture is to recognize entities and add the new entities into the semantic lexicon. This can be achieved through the initial phase of the bootstrapping process which selects a subset of the extraction patterns that tend to extract the seed words. This is referred as the *pattern pool* in [7]. The nouns extracted by these patterns become candidates for the lexicon and are placed in a *candidate word pool*. Basilisk scores each candidate word, and the five best candidate words are added to the lexicon and the process will continue until the rest of the document is processed. The overall Basilisk Algorithm is shown in Fig. 3.

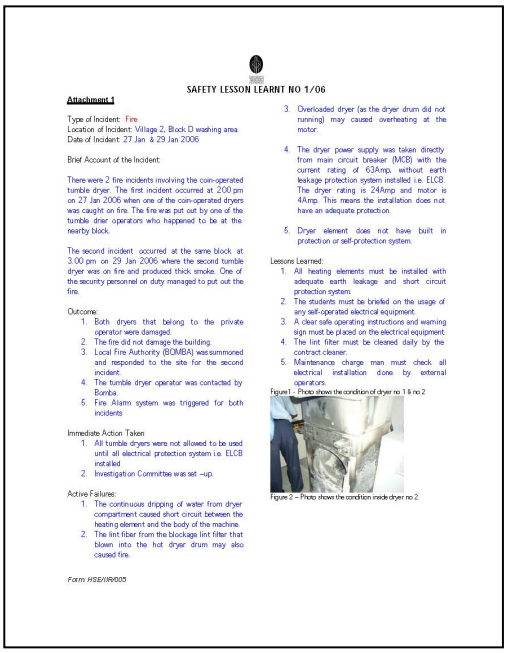
Generate all extraction patterns in the corpus and record their extractions.
lexicon = {seed words}
$i := 0$

BOOTSTRAPPING
1. Score all extraction patterns
2. *pattern pool* = top ranked 20+i patterns
3. *candidate word pool* = extractions of patterns in *pattern pool*
4. Score candidate words in *candidate word pool*
5. Add top 5 candidate words to *lexicon*
6. $i := i + 1$
7. Go to Step 1.

Figure 3. Basilisk Bootstrapping Algorithm

V. CONCEPTUAL WORKING EXAMPLE

Input for the system is accidental report document that contains not only texts but also images. Fig. 4 shows an example of the document.



SAFETY LESSON LEARNT NO 1/06

Attachment 1

Type of Incident: Fire
Location of Incident: Village 2, Block D washing area.
Date of Incident: 27 Jan & 29 Jan 2006

Brief Account of the Incident:

There were 2 fire incidents involving the coin-operated tumble dryer. The first incident occurred at 2.00 pm on 27 Jan 2006 when one of the coin-operated dryers was caught on fire. The fire was put out by one of the tumble drier operators who happened to be at the nearby block.

The second incident occurred at the same block at 3.00 pm on 29 Jan 2006 where the second tumble dryer was on fire and produced thick smoke. One of the security personnel on duty managed to put out the fire.

Outcome:
1. Both dryers that belong to the private operator were damaged.
2. The fire did not damage the building.
3. Local Fire Authority (BOMBA) was summoned and responded to the site for the second incident.
4. The tumble dryer operator was contacted by Bomba.
5. Fire Alarm system was triggered for both incidents

Immediate Action Taken
1. All tumble dryers were not allowed to be used until all electrical protection system i.e. ELCB installed.
2. Investigation Committee was set –up.

Active Failures:
1. The continuous dripping of water from dryer compartment caused short circuit between the heating element and the body of the machine.
2. The lint filter from the blockage lint filter that blown into the hot dryer drum may also caused fire.

Form: HSE/IR/005

3. Overloaded dryer (as the dryer drum did not running) may caused overheating at the motor.
4. The dryer power supply was taken directly from main circuit breaker (MCB) with the current rating of 63Amp, without earth leakage protection system installed i.e. ELCB. The dryer rating is 24Amp and motor is 4Amp. This means the installation does not have an adequate protection.
6. Dryer element does not have built in protection or self-protection system.

Lessons Learned:
1. All heating elements must be installed with adequate earth leakage and short circuit protection system.
2. The students must be briefed on the usage of any self-operated electrical equipment.
3. A clear safe operating instructions and warning sign must be placed on the electrical equipment.
4. The lint filter must be cleaned daily by the contract cleaner.
5. Maintenance charge man must check all electrical installation done by external operators.

Figure 1 - Photo shows the condition of dryer no 1 & no 2

Figure 2 – Photo shows the condition inside dryer no 2.

Figure 4. UTP's HSE Accident Report

Process 1 will remove the image and other noisy elements. Then, clean text will be parsed into sentences. In this process, LG parses the sentences into words and create a linkage between words that have association. For instance from the report we will identify accident date. For example as illustrated in Fig. 5, LG create links and assign it by specific connector.

In Process 2, new extraction pattern is created by considering LG link. Assume that in the initial pattern, we have patterns like the incident happened on <date>, the accident occurred at <time>, etc. Example of new extraction pattern will be like the incident happened ON <TM+ TY>. New extraction pattern is created by considering ON, TM and TY connector to identify date entity. ON is used to connect preposition "on" to certain time expression, TM is used to connect month names to day numbers and TY is used for certain idiomatic usages of year numbers. The process then proceeds to the next stage, which is the application of Basilisk Algorithm.

In this process, first all new extraction pattern will be assigned a score. And the top ("n" + i) patterns will be stored in the extraction pattern pool. Where the value of "n" is determined by the user. Applying those extraction patterns into the corpus will result in the derivation of candidate word. Best candidate will be tagged with entity attribute and "n" best candidates are stored into the semantic lexicon.

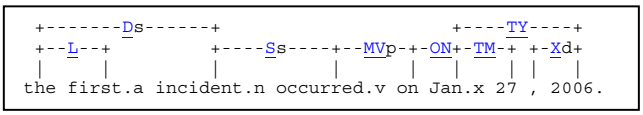


Figure 5. Create Extraction Pattern using LG Parser

VI. CONCLUSION

This paper presents a hybrid approach which combines the LG and Basilisk Algorithm to improve the NER system demonstrated on the UTP's HSE reports. With semi-supervised learning approach, we hope that our algorithm will help to create the semantic lexicon and thus improve the efficiency and accuracy of the system.

REFERENCES

[1] R. Grishman and B. Sundheim "Message Understanding Conference -6 : A Brief History " in 16th Conference on Computational Linguistics, Copenhagen, Denmark , pp. 466-471.

[2] D. Nadeau, "Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision " in Ottawa-Carleton Institute for Computer Science School of Information Technology and Engineering, Canada: University of Ottawa 2007.

[3] K. Shaalan and H. Raza, "Person Name Entity Recognition for Arabic " in 5th Workshop on Important Unresolved Matters, Prague, Czech Republic, 2007, pp. 17–24.

[4] D. Maynard, K. Bontcheva, H. Saggion, H. Cunningham, and O. Hamza, "Using a Text Engineering Framework to Build an Extendable and Portable IE-based Summarisation System," ACL Workshop on Text Summarisation, Philadelphia, 2002.

[5] H. Cunningham, "Software Architecture for Language Engineering," in Department of Computer Science: University of Sheffield, 2000.

[6] D. D. Sleator and D. Temperley, "Parsing English with Link Grammars," in 3rd International Workshop on Parsing Technologies (ACL – SIGPARSE), University of Tilburg, Netherlands, 1993.

[7] M. Thelen and E. Rillof, "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts," in Conference on Empirical Methods in Natural Language Processing, 2002, pp. 214-221.

[8] G. S. Mann and D. Yarowsky, "Unsupervised personal name disambiguation," in the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, Canada, 2003, pp. 33-40.

[9] D. Nadeau and P. D. Turney, "A Supervised Learning Approach to Acronym Identification," in 18th Conference of the Canadian Society for Computational Studies of Intelligence, Victoria, BC, Canada, 2005, pp. 319–329.

[10] G. Paliouras, V. Karkaletsis, G. Petasis, and C. D. Spyropoulos, "Learning Decision Trees for Named-Entity Recognition and Classification," in 14th European Conference on Artificial Intelligence (ECAI 2000), Berlin, Germany, 2000.

[11] E. Rillof, "Automatically Generating Extraction Patterns from Untagged Tex," in Thirteenth National Conference on Articial Intelligence (AAAI-96), 1996, pp. 1044-1049.

[12] A. Mansouri, L. Suriani Affendey, and A. Mamat, "Name Entity Recognition Approach," International Journal of Computer Science and Network Security, vol. 8, 2008.

[13] R. Yangarber, W. Lin, and R. Grishman, "Unsupervised Learning of Generalized Names," in 19th International Conference on Computational Linguistic (COLING 2002), Taipei, 2002, pp. 1-7.

[14] H. V. Madhyastha, N. Balakrishnan, and K. R. Ramakrishnan, "Event Information Extraction Using Link Grammar," in 13th International Workshop on Multi-lingual Information Management, 2003, pp. 16-22.